CS534 - Machine Learning

Written Homework Assignment 1

Author: Vy Bui

OSUID: 934370552

Email: buivy@oregonstate.edu

The School of Electrical Engineering and Computer Science

Oregon State University

Problem 1
[10pts]

(a) The log likelihood function of $\mathbf{w}$ is

$$l(w) = log \prod_{i=1}^{N} P(y_i|\mathbf{x_i};\mathbf{w})$$

(b)
First, maximizing the log likelihood function is equivalent to minimizing the negative log likelihood function.

$$\arg\max_w log(w) = \arg\min_w -log(w) = \arg\min_w -log \prod_{i=1}^{N} P(y_i|\mathbf{x_i};\mathbf{w}) = \arg\min_w -\sum_{i=1}^{N} logp(y_i|\mathbf{x_i};\mathbf{w}) \quad (1)$$

Because the likelyhood is Gaussian,

$$logp(y_i|\mathbf{x_i};\mathbf{w}) = -\frac{(y_i - \mathbf{x_i^T w})^2}{2\sigma^2} + const \qquad (2)$$

where the const consists of all terms independent of $\mathbf{w}$.
Substituting (2) to (1) and droping const results in

$$\arg\max_w log(w) = \arg\min_w -\sum_{i=1}^{N} logp(y_i|\mathbf{x_i};\mathbf{w}) = \arg\min_w \sum_{i=1}^{N} \frac{(y_i - \mathbf{x_i^T w})^2}{2\sigma^2}$$

$$= \frac{1}{2}\sum_{i=1}^{N} a_i(\mathbf{w^T x_i} - y_i)^2, a_i = \frac{1}{\sigma^2} \qquad (3)$$

(c) The gradient of $L(\mathbf{w})$ can be computed as following

$$\nabla L(w) = \frac{1}{\sigma^2}\sum_{i=1}^{N}(\mathbf{w^T x_i} - y_i)\mathbf{x_i}$$

Because the gradient points in the direction of steepest ascent, we need to go in the opposite direction to reach the minimum, one step at a time. Hence, the update rule is

$$\mathbf{w} \leftarrow \mathbf{w} - \gamma\nabla L(w) = \mathbf{w} - \frac{1}{\sigma^2}\sum_{i=1}^{N}(\mathbf{w^T x_i} - y_i)\mathbf{x_i}$$

(d) Because (3) is a quadratic function of **w**, we can compute the global optimum by setting its gradient to 0 and solve for **w**.

$$\frac{l}{\partial \mathbf{w}} = \frac{1}{2\sigma^2}\frac{d}{d\mathbf{w}}((\mathbf{y}-\mathbf{xw})^{\mathbf{T}}(\mathbf{y}-\mathbf{xw})) = \frac{1}{2\sigma^2}\frac{d}{d\mathbf{w}}((\mathbf{y^Ty}-\mathbf{2y^Txw}+\mathbf{w^Tx^Txw}))$$

$$= \frac{1}{2\sigma^2}\frac{d}{d\mathbf{w}}((-\mathbf{y^TX}+\mathbf{w^Tx^Tx})) \qquad (4)$$

Setting (4) to 0 results in

$$\mathbf{w_{op}^T x^T x} = \mathbf{y^T X} \Leftrightarrow \mathbf{w_{op}^T} = \mathbf{y^T x}(\mathbf{x^T x})^{\mathbf{-1}} \Leftrightarrow \mathbf{w_{op}} = (\mathbf{x^T x})^{\mathbf{-1}}\mathbf{x^T y}$$

Problem 2
[10pts]

(a) Compute the log-likelihood function

$$l(w) = logL(w) = \sum_{i=1}^{N}\sum_{k=1}^{K} log(p(y=k|x_i)^{y_{ik}}) = \sum_{i=1}^{N}\sum_{k=1}^{K} log(\frac{e^{\mathbf{w_k^T x_i}}}{\sum_{j=1}^{K} e^{\mathbf{w_j^T x_i}}})^{y_{ik}}$$

(b) Compute the gradient of the log-likelihood function with regard to the weight vector $\mathbf{w}_c$ of class c.

First, the function can be simplified as follows

$$l(w) = \sum_{i=1}^{N}\sum_{k=1}^{K} log(\frac{e^{\mathbf{w_k^T x_i}}}{\sum_{j=1}^{K} e^{\mathbf{w_j^T x_i}}})^{y_{ik}} = \sum_{i=1}^{N}\sum_{k=1}^{K} y_{ik} log(\frac{e^{\mathbf{w_k^T x_i}}}{\sum_{j=1}^{K} e^{\mathbf{w_j^T x_i}}})$$

$$= \sum_{i=1}^{N}\sum_{k=1}^{K} y_{ik}(\mathbf{w_k^T x_i} - log\sum_{j=1}^{K} e^{\mathbf{w_j^T x_i}}) = \sum_{i=1}^{N}(\sum_{k=1}^{K} y_{ik}\mathbf{w_k^T x_i} - \sum_{k=1}^{K}(y_{ik} log\sum_{j=1}^{K} e^{\mathbf{w_j^T x_i}}))$$

$$= \sum_{i=1}^{N}(\sum_{k=1}^{K} y_{ik}\mathbf{w_k^T x_i} - log\sum_{j=1}^{K} e^{\mathbf{w_j^T x_i}})$$

Now it is easier to calculate the gradient as follows

$$\nabla_{w_{ch}} l(w) = \frac{\partial}{\partial w_{ch}}(\sum_{i=1}^{N}(\sum_{k=1}^{K} y_{ik}\mathbf{w_k^T x_i} - log\sum_{j=1}^{K} e^{\mathbf{w_j^T x_i}})) = \sum_{i=1}^{N}(y_{ic}x_i^h - \frac{x_i^h e^{\mathbf{w_c^T x_i}}}{\sum_{j=1}^{K} e^{\mathbf{w_j^T x_i}}})$$

$$= \sum_{i=1}^{N} x_i^h(y_{ic} - \frac{e^{\mathbf{w_c^T x_i}}}{\sum_{j=1}^{K} e^{\mathbf{w_j^T x_i}}})$$

where $h = 0, 1, 2, ..., n$ and $c = 0, 1, 2, ..., k$.

Therefore, the gradient with regard to vector w_c is

$$\nabla_{w_c} l(w) = \sum_{i=1}^{N} x_i(y_{ic} - P(y_i = c|x_i; w))$$

Problem 3
[10pts]

(a) Derive the posterior distribution $p(\hat{\theta}|X_1, ..., X_n, \alpha, \beta)$ and show that it is also a Beta distribution.

First, the likelihood function can be computed as follows

$$p(X_1, ..., X_n|\hat{\theta}) = \prod_{i=1}^{n} \hat{\theta}^{x_i}(1-\hat{\theta})^{1-x_i} = \hat{\theta}^{\sum x_i}(1-\hat{\theta})^{n-\sum x_i}$$

Then, the posterior of $\hat{\theta}|X_1, ..., X_n$ is

$$p(\hat{\theta}|X_1, ..., X_n, \alpha, \beta) \propto p(X_1, ..., X_n, \alpha, \beta|\hat{\theta})p(\hat{\theta})$$

$$= \hat{\theta}^{\sum x_i}(1-\hat{\theta})^{n-\sum x_i} \frac{\hat{\theta}^{\alpha-1}(1-\hat{\theta})^{\beta-1}}{B(\alpha, \beta)} = \frac{\hat{\theta}^{\alpha+\sum x_i}(1-\hat{\theta})^{\beta+n-\sum x_i}}{B(\alpha, \beta)}$$

$$= Beta(\hat{\theta}|\alpha + \sum x_i, \beta + n - \sum x_i)$$

(b) For the case of observing 2 heads out of 5 tosses, the posterior distribution of θ can be calculated as follows

$$Beta(\theta|2 + 2, 2 + 5 - 2) = Beta(\theta|4, 5)$$

That for the case of observing 20 heads out of 50 tosses is

$$Beta(\theta|2 + 20, 2 + 50 - 20) = Beta(\theta|22, 32)$$

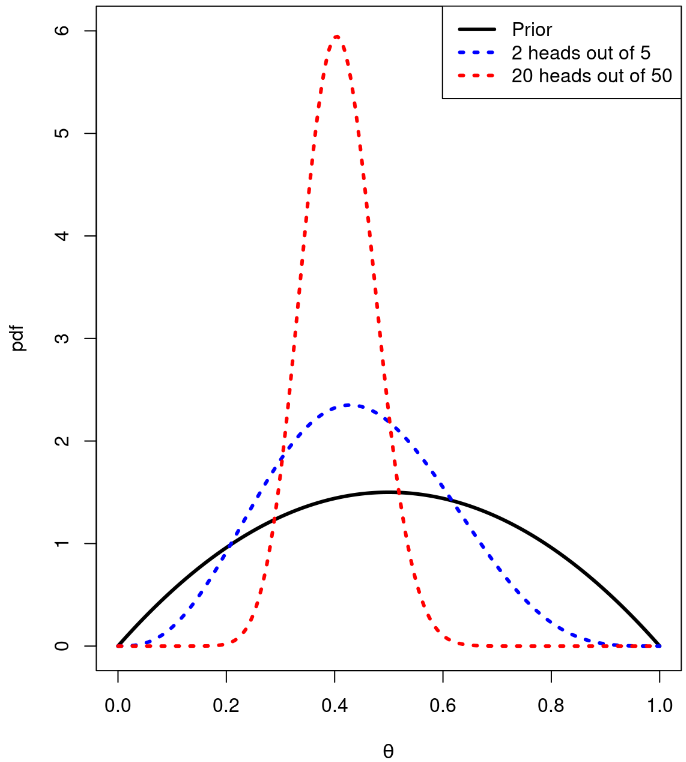

Figure 1: The posterior distributions of θ

The posterior distribution's spread will get smaller and smaller, centered at $\theta = 0.4$.